Does typing or handwriting exam responses make any difference? Evidence from the literature

Santi Lestari (Research Division)

Introduction

Computer-based tests have become widespread in many assessment contexts, including language assessments and university admissions tests. Many general qualifications exams, however, remain in a paper-based mode, often requiring students to handwrite long answers, such as essays, under time constraints. Insufficient and unequal digital provision across schools is often identified as a major barrier to a full adoption of computer-based exams for general qualifications in many jurisdictions, including in England (Coombe et al., 2020). One feasible approach to overcoming this barrier is a gradual adoption, which involves offering both modes of exam administration in parallel (i.e., paper-based and computer-based) (Arce-Ferrer & Bulut, 2018; Coombe et al., 2020). This approach, however, presents risks of mode effects (Coombe et al., 2020). Mode effects occur when there are unavoidable differences between paper-based and computer-based exams that are intended to be equivalent. This can mean the exams measure slightly different constructs and the resulting scores may not be directly equivalent. When an exam is offered in both paper-based and computer-based modes, and results from both are treated as equivalent, and therefore interchangeable, the comparability between modes needs to be ascertained. This includes investigating potential response mode effects for extended writing questions, or, in other words, examining whether the mode in which students respond to the questions (i.e., by handwriting or typing on the computer) introduces systematic differences. We conducted a literature review on writing response mode effects, and this article summarises the key findings.

Methods

To identify the relevant studies, we searched major databases in education, psychology and linguistics, including Education Resources Information Center (ERIC) and Linguistics and Language Behavior Abstracts (LLBA). Keywords used in the searches included words related to i) writing mode such as “typed”, “typing”, “word-processed”, “handwritten” and “handwriting”, and ii) assessment such as “exam”, “examination”, “test” and “exam script”. We also checked the reference lists of the selected studies to find additional studies.

© Cambridge University Press & Assessment 2024

The criteria for inclusion in the review were that studies had to: a) be published in English; b) compare the two writing modes (i.e., handwriting and typing/word processing) in an assessment context; c) involve an assessment that required an extended writing response; and d) involve empirical data (i.e., using students' writing performance data from either an operational exam administration and/or an experimental setting). We decided to include various publication types (i.e., peer-reviewed journal articles, conference papers, doctoral theses/dissertations and institutional reports). This is because research investigating mode effects, especially for high-stakes assessments, is often conducted by awarding organisations and published only as an institutional report. We read the selected studies to identify the research context, focus and key findings.

Findings

Overview of the studies included

A total of 47 studies, published between 1990 and 2021, were included in the review (Figure 1). These studies varied in terms of context and focus. Figure 2 summarises the number of studies by research context. Almost half of the studies (22 out of 47) were conducted in language assessment contexts, almost exclusively in English as a second or foreign language (ESL/EFL) contexts (e.g., Brunfaut et al., 2018; Chan et al., 2018; Lessien, 2013; Manalo & Wolfe, 2000a), with only one study investigating mode effects in another language, namely Mandarin Chinese as a foreign language (Zhu et al., 2016). It is not surprising that language assessment is the dominant context given that writing as part of language proficiency is commonly tested in direct language assessments.[1] Some of the ESL/EFL assessments are also high-stakes in nature because important, often life-changing, decisions are made based on the test scores, giving more reason to investigate potential mode effects.

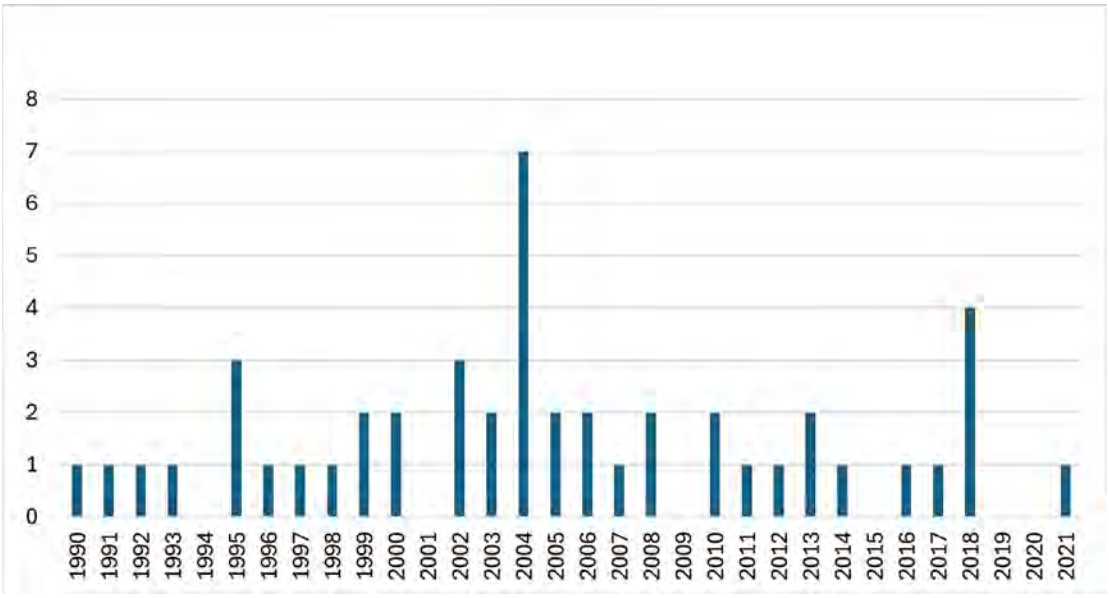

Figure 1: Number of studies across the years (n=47)

1 As opposed to indirect language assessments which measure writing proficiency through means other than directly requiring candidates to write, e.g., error recognition.

© Cambridge University Press & Assessment 2024

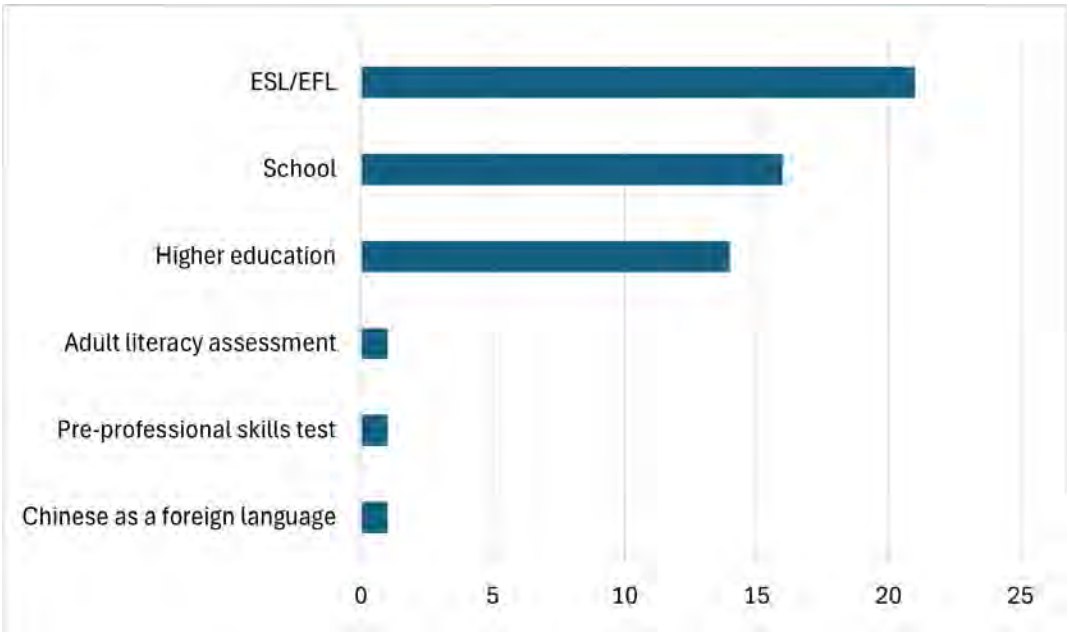

Figure 2: Number of studies by research context (n=47, multiple contexts possible)

In terms of level of education, 16 studies were conducted in school contexts, the majority of which were in the US (e.g., Burke & Cizek, 2006; Hollenbeck et al., 1999; Russell & Tao, 2004b; Wolfe et al., 1995; Wolfe et al., 1996). There are also school-based studies investigating mode effects in other jurisdictions including the UK (Charman, 2014; Connelly et al., 2007), Australia (MacCann et al., 2002) and Hong Kong (Lam & Pennington, 1995). Fourteen studies were conducted in higher education contexts. Such studies might focus on ESL/EFL (e.g., Jin & Yan, 2017; Kim et al., 2018), a non-language subject, such as theology (e.g., Mogey & Hartley, 2013) or admissions tests (e.g., Bridgeman & Cooper, 1998). Two studies do not fit into these education levels: Chen et al. (2011) studied mode effects of adult literacy assessment in the US, called the National Assessment of Adult Literacy (NAAL), and Yu et al. (2004) examined mode effects of the essay writing component of the Praxis Pre-Professional Skills Test, a battery test assessing basic academic skills of pre-service teachers.

Studies also varied in terms of the focus of their investigation (Figure 3). The primary focus of most studies was on the comparability of students' performance across the two modes of writing. Most studies operationalised performance as scores (e.g., Lam & Pennington, 1995; Yu & Iwashita, 2021), but some also examined the comparability of the characteristics of the texts produced (e.g., Barkaoui & Knouzi, 2018; Chambers, 2008; Charman, 2014; Jin & Yan, 2017) and a few investigated the comparability of students' composing processes across the two modes (Chan et al., 2018; Jin & Yan, 2017; Lee, 2002; Wolfe et al., 1993).

Researchers examining the comparability of scores across the two writing modes also often gathered students' contextual information, including demographic data such as gender, ethnicity and socio-economic background (e.g., Bridgeman & Cooper, 1998; Chen et al., 2011), language proficiency level (e.g., Lessien, 2013;

© Cambridge University Press & Assessment 2024

Manalo & Wolfe, 2000a) and information on students' computer familiarity[2] and/or perceptions of the composition mode (e.g., Barkaoui & Knouzi, 2018; Jin & Yan, 2017; Whithaus et al., 2008; Wolfe et al., 1996). Contextual information is useful to allow more fine-grained analyses of writing mode effects across sub-groups of a candidate population or to explain the presence of mode effects, if any.

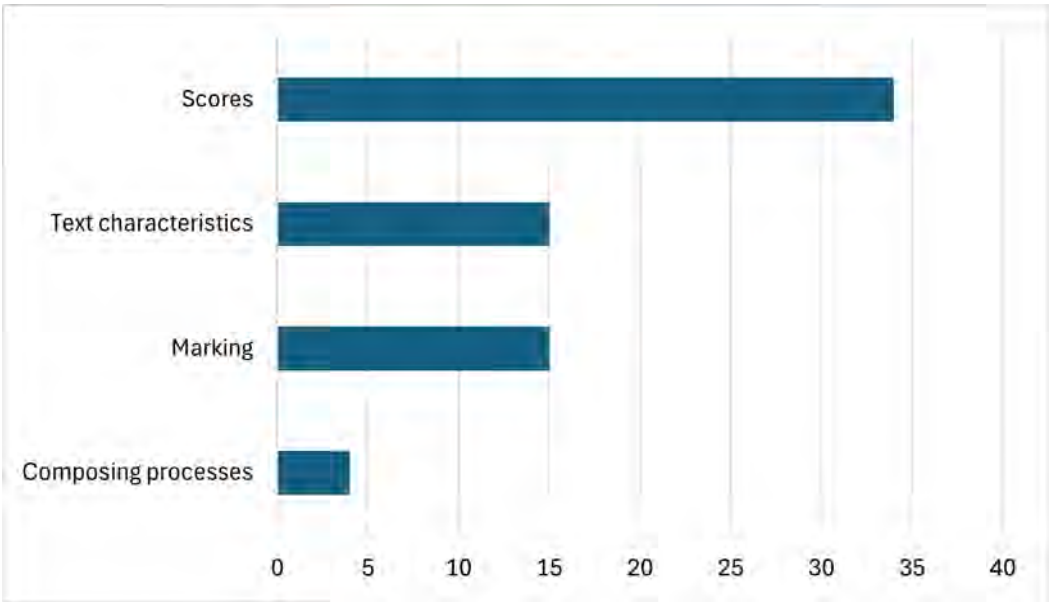

Figure 3: Number of studies by research focus (n=47, multiple focuses possible)

Another research focus was the effect of presentation mode on essay marking. More precisely, the research typically looked into whether the mode of script presentation to markers (i.e., handwritten or typed/word-processed) had differential effects on marking quality including marker bias (e.g., Arnold et al., 1990; Brown, 2003; Chen et al., 2011), marking processes (Wolfe et al., 1993), and other measures of marking quality such as inter-rater agreement and/or reliability (Lee, 2004; Manalo & Wolfe, 2000b).

The following sections present key findings under each research focus.

Comparability of scores

As the research methods used to investigate the comparability of scores vary considerably, it is important to be cautious in drawing conclusions from different research findings. Arce-Ferrer and Bulut (2018) examined four commonly used data collection designs[3] in mode effects studies and concluded that the single-group design with counterbalancing and random-groups design were the superior data collection designs at detecting mode effects at the test level (i.e., score distributions). Furthermore, if a score comparison is made at the group level, rather than the individual level, the score comparability conclusion may also hold true at the group level only. It is typically the case with the studies included in this

2 The term "computer familiarity" is used in the current article to include typing skills, word processing skills, experience or frequency of using a computer and level of comfort or confidence in using a computer.
3 Single-group design with counterbalancing, single-group design without counterbalancing, random-groups design, and anchor-test-nonequivalent-groups design.

© Cambridge University Press & Assessment 2024

review that the score comparability analysis was conducted at the group level rather than individual level, for example, by comparing the mean scores of each group.

Typed essays scored higher than handwritten essays

Some studies found that students performed better when they typed or word-processed their essay than when they handwrote it (Lam & Pennington, 1995; Lessien, 2013; Russell & Haney, 1997; Russell & Plati, 2002; Zhu et al., 2016). Russell and Haney (1997) and Lessien (2013) also found that the writing mode effect was highly significant, favouring typed essays, and this was particularly the case for students with high proficiency in English (Lessien, 2013). Findings from Zhu et al. (2016) were particularly interesting as this study investigated writing mode effects in Mandarin Chinese as a foreign language. Most of the students in the study also reported that they preferred word-processing their essay to handwriting it, as they felt word-processed essays appeared more professional.

Typed essays scored lower than handwritten essays

Other studies found that students performed better when they handwrote their essay than when they typed it (Breland et al., 2004; Bridgeman & Cooper, 1998; Chen et al., 2011; Connelly et al., 2007; Green & Maycock, 2004; Manalo & Wolfe, 2000a; McGuire, 1995; Yu et al., 2004). Manalo and Wolfe (2000a) found that when language proficiency was controlled for, the handwritten essays were scored approximately one-third of a standard deviation higher than the typed essays. Researching writing mode effects among primary school students aged 4 to 11 years old, Connelly et al. (2007) found that the quality of the handwritten scripts was better than that of the typed scripts. A differential effect of writing mode was also observed in Chen et al. (2011), whereby the computer-based mode disadvantaged unemployed candidates even more than employed candidates. Bridgeman and Cooper (1998), using the Graduate Management Admission Test (GMAT) essay task, observed that the score difference favouring handwriting mode did not interact with candidates' gender, ethnicity or English as a second language group classification.

No meaningful score difference between typed and handwritten essays

Additional studies found that generally there were no (meaningful) writing mode effects on students' performance (Barkaoui & Knouzi, 2018; Brunfaut et al., 2018; Chan et al., 2018; Charman, 2014; Horkay et al., 2006; Lee, 2002; Lovett et al., 2010; MacCann et al., 2002; Mogey et al., 2010; Yu & Iwashita, 2021). For instance, Chan et al. (2018), investigating the comparability of paper-based and computer-based delivery of the IELTS Writing test, found that scores across both modes were generally comparable although candidates scored better in the Lexical Resources criterion when they handwrote their essay. Chan et al. (2018) theorise that different writing modes might elicit certain aspects of writing, in this case lexical resources, slightly differently. Furthermore, they also observed that some aspects of computer familiarity significantly predicted performance in computer-based writing assessment, confirming findings from an earlier study by Horkay et al. (2006).

© Cambridge University Press & Assessment 2024

Mode effects and contextual variables

Writing mode effects are not always straightforward and can be influenced by the students' contextual factors and methods used in scoring the writing. Students' computer familiarity and typing speed were found to interact with writing mode (Russell, 1999; Wolfe et al., 1995). Students with greater familiarity with word processing software tended to perform equally well either typing or handwriting their essay, whereas students with less word processing experience tended to perform better and write more when handwriting their essay (Wolfe et al., 1995). Students' language proficiency is another factor that may influence writing mode effects. Students with weaker English language ability tended to perform better on handwritten essays, while those with better English performed comparably on both writing modes (Wolfe & Manalo, 2004). A similar finding was also observed by Brunfaut et al. (2018) in that the student group taking the lowest level of the English proficiency tests found a writing task easier in the handwriting mode than in the typing mode. Scoring method (i.e., holistic versus analytic) was also found to influence the scores of writing produced under the two writing modes. When holistic rating was used, no significant mean score difference was observed across the two modes; however, word-processed essays received significantly higher scores when analytic scoring was used (Lee, 2004).

Comparability of marking

The focus of marking comparability is on the effect of essay presentation mode on the marker (i.e., whether markers give different scores to the handwritten and typed versions of the same essays). Marker bias (i.e., whether markers give systematically higher scores on one presentation mode over another) was the primary focus of most studies examining comparability of marking across the two presentation modes. A few studies, however, also focused on the comparability of inter-rater agreement and reliability across the two modes. Some studies examined markers' perceptions of scoring essays in the two modes.

Marker bias

Handwritten essays were generally found to receive higher scores than the typed or word-processed versions of the same essays (Arnold et al., 1990; Brown, 2003; MacCann et al., 2002; Powers et al., 1994; Russell & Tao, 2004a; Shaw, 2003; Sweedler-Brown, 1991). The magnitude of the marker bias sometimes varied across different levels of performance. For example, Sweedler-Brown (1991) found that marking bias was more prominent for higher level performance; there was a significant difference in scores between modes for essays that received higher scores in the original handwritten format, but not for essays that received lower scores in the original handwritten format. Brown (2003) also found that the bias effect was moderated by the legibility of the handwriting, in that the score difference was higher for essays with poor legibility. This suggests that students with poor handwriting were, surprisingly, somewhat advantaged.

Chen et al. (2011), conversely, found no statistically or practically significant difference in the scores awarded to the typed and handwritten versions of essays. Similarly, Green and Maycock (2004) found that presentation mode effect was only negligible and of no practical importance.

© Cambridge University Press & Assessment 2024

Several potential explanations were identified for the common finding of bias against typed essays. Markers tended to have a higher expectation of word-processed essays (Arnold et al., 1990; Russell & Tao, 2004a). Word-processed essays were also often perceived to be shorter than handwritten essays although they were exactly of the same length (Arnold et al., 1990; Powers et al., 1994). Altering formatting style such as space and font size to make the word-processed essays appear to have a similar length to the handwritten version was found to reduce the size of presentation mode effect in Powers et al. (1994) but not in Russell and Tao (2004a).

Although word-processed essays were found to be easier to read, surface errors such as spelling and punctuation errors tended to appear more prominent and therefore more recognisable (Arnold et al., 1990; Russell & Tao, 2004a; Shaw, 2003; Wolfe et al., 1993). Handwriting, especially poor handwriting, could also mask such errors (Powers et al., 1994), which might explain Brown's (2003) finding above. Some markers in Russell and Tao (2004a) also reported that they could see students' effort more in handwritten essays, echoing findings from Powers et al. (1994) suggesting that traces of revisions in handwritten essays, such as strikethroughs, seemed to be valued by markers (who were usually also teachers). These factors may explain the bias against word-processed essays.

Marking reliability

Markers were generally found to have stronger agreement when scoring essays in the word-processed format than in the handwritten format. For example, Lee (2004) found that markers reached higher percentages of exact agreement when marking word-processed (76.1 per cent) and transcribed essays (78.6 per cent) than when marking handwritten essays (64.3 per cent). Furthermore, using other measures of inter-rater agreement and reliability (i.e., Pearson product moment correlation and Cohen's kappa), Manalo and Wolfe (2000b) and Wolfe and Manalo (2005) found that it was easier for markers to agree on scores for the word-processed essays than for the handwritten ones. Markers in Shaw (2003) reported that word-processed essays had a more similar general appearance and that both strong and weak essays were easier to read, potentially contributing to the increased objectivity.

Differences in scoring processes

The analysis of think-aloud protocol data in Wolfe et al. (1993) revealed differences in the processes involved in marking handwritten and word-processed scripts. When reading the handwritten essays, markers read less at a time and paused more often to make evaluative comments about the essay. In contrast, when reading the word-processed essays, they paused less frequently and saved most of the comments until after finishing reading the entire essay. Commentary on the word-processed essays tended to focus on the development of the essay, while comments on the handwritten essays focused more on essay organisation and authorial voice.

Comparability of text characteristics

Text length, typically measured in word and/or sentence count, is the most common measure of text characteristics explored in the studies that were

© Cambridge University Press & Assessment 2024

reviewed. Students tended to write longer texts when using a computer than when writing by hand (e.g., Barkaoui & Knouzi, 2018; Jin & Yan, 2017; Kim et al., 2018; Lee, 2002; Lovett et al., 2010; Mogey et al., 2010; Russell & Haney, 1997). However, this difference was not always statistically significant. The use of the keyboard could potentially explain the increased fluency in computer-based writing tests (Kim et al., 2018). Some studies found that text length also varied more considerably in word-processed essays than in handwritten ones (e.g., Chen et al., 2011; Endres, 2012).

In terms of language complexity, word-processed essays were found to have higher lexical variation (Barkaoui & Knouzi, 2018; Chambers, 2008; Charman, 2014), and more sophisticated vocabulary and varied syntactic structures (Barkaoui & Knouzi, 2018). Kim et al. (2018) also found similar patterns of results but commented that the differences were unlikely to be meaningful as average differences were relatively small and there was considerable overlap in values between the two modes.

Errors, usually mechanical errors such as punctuation and capitalisation, were also an area of investigation under text characteristics. It was generally found that there were no major differences in terms of the frequency of errors in handwritten and word-processed essays (e.g., Chambers, 2008; Endres, 2012; Wolfe et al., 1996). However, the nature of errors might differ. For example, Endres (2012) found that spelling errors in computer-based English writing tests were mainly typographical errors, which were potentially caused by typing errors, whereas spelling errors in the equivalent paper-based tests tended to be more developmental errors, potentially resulting from first language interference. Jin and Yan (2017), however, found that students made significantly fewer errors when they typed their essays than when they handwrote them, even though editing tools, such as grammar- and spell-checkers, were disabled.

Other features of text characteristics examined in previous studies include tone and readability. Whithaus et al. (2008) found that informal tone was perceived to be less present in typed essays than in handwritten ones. Using various readability indices including Flesch Reading Ease scores and Fog index, Mogey and Hartley (2013) found that the typed essays were generally more readable than the handwritten ones.

Most studies examining the comparability of text characteristics, however, did not consider students' level of computer familiarity. Including this aspect in their study, Wolfe et al. (1996) found that using a word processor did not impact the writing quality of students with medium and high levels of computer familiarity, but it harshly impacted those with lower levels of computer familiarity. On text length, specifically, students with medium and high levels of computer familiarity wrote longer word-processed essays than handwritten essays. In contrast, students with low familiarity wrote over 100 words fewer on average on a word processor than on paper. Furthermore, students with a medium or high level of computer familiarity tended to write a higher number of simple sentences when handwriting their essays compared to when typing them. Conversely, those with a low level

© Cambridge University Press & Assessment 2024

of computer familiarity tended to write more simple sentences when typing compared to when handwriting their essays.

In summary, differences in terms of text characteristics were observed between typed and handwritten essays. These differences, however, were not always statistically significant and/or of practical importance, and, furthermore, were not necessarily reflected in scores (Barkaoui & Knouzi, 2018). It should also be noted that writing modes might have differential effects on students with different levels of computer familiarity, as Wolfe et al. (1996) observed.

Comparability of composing processes

Composing processes refer to the activities that students engage in when answering an extended writing question. Chan et al. (2018) and Jin & Yan (2017) found that both writing modes elicited similar composing processes. However, a few differences were observed. In Jin and Yan's (2017) study, students with low and moderate levels of computer familiarity admitted that they planned better when handwriting their essay in the paper-based mode. One candidate explained that as they were required to handwrite their essay using a pen in the paper-based mode, they were more inclined to plan more carefully before writing to avoid making many corrections during writing, which would affect the essay presentation. In contrast, typing their essay on the computer allowed them to review and edit their essay more flexibly and therefore they were less inclined to plan more carefully before writing (Jin & Yan, 2017). Similarly, Chan et al. (2018) found some minor differences especially in planning, generating texts and monitoring and revising, although these differences in composing processes might not necessarily be reflected in scores. In terms of revising, some students in the study reported that when handwriting their essay in the paper-based mode, they tended to focus more on word level revisions, but when typing their essay in the computer-based mode, they tended to revise at the clause and sentence levels. Again, these differences were likely to be due to the flexibility afforded by the computer-based mode.

Discussion and conclusion

The question of whether typing or handwriting answers to extended writing questions in exams makes a difference has been widely investigated although the context and focus on which research has been conducted varied. In terms of context, more studies have been carried out in the context of English as a second or foreign language assessment, including proficiency and placement tests in higher education settings. Studies in the context of school education have been conducted in the US more than in any other jurisdiction, although this could be due to publication bias as we selected only articles and reports published in the English language. In terms of research focus, four aspects of comparability have been investigated: scores, marking, text characteristics and composing processes.

For **comparability of scores**, we could see that more studies, particularly the recent ones (which often used more robust methods involving the single-group design with counterbalancing and controlling for contextual factors), tended to find that scores across the two writing modes were comparable, at least at

© Cambridge University Press & Assessment 2024

the group level. However, there were also non-trivial numbers of studies that found a mode effect in one direction or the other. In a few studies, two contextual factors have been found to interact with mode effects: English proficiency and computer familiarity. Students with weaker English language ability tended to perform better on handwritten essays, while those with better English performed comparably on both modes. This particularly concerns writing mode effects in the context of ESL/EFL assessments. One implication is that when designing tests targeted specifically at students with low language proficiency, test designers may need to carefully consider whether to require students to type their essay as typing may underestimate the measurement of their writing ability.

Students with greater familiarity with word processing software tended to perform equally well either typing or handwriting their essay, whereas students with less experience with word processing tended to perform better and write more when handwriting their essay. It is therefore important to ensure that students have a sufficient level of computer familiarity, especially typing and word processing skills, to perform the assessment tasks. When it is known that a candidate pool varies considerably in their level of computer familiarity, it is recommended for test developers to offer both options of writing mode. However, as computer literacy is considered an indispensable aspect of academic literacy in the 2Ist century, some may argue that computer literacy should be considered an important element of the construct measured both in language assessment and in the assessment of other subjects (see e.g., Jin & Yan, 2017).

In terms of **comparability of marking**, handwritten essays generally appeared to receive higher scores than the word-processed version of the same essays. Reasons for this include markers having a higher expectation of word-processed essays and that word-processed essays were often perceived to be shorter than the handwritten version. As word-processed essays are easier to read, surface and mechanical errors such as spelling and punctuation become more recognisable to markers. On the other hand, handwriting, especially with low legibility, could mask such errors. Markers (who are usually teachers) also seemed to appreciate traces of corrections in handwritten essays such as strikethroughs, further contributing to bias against typed essays.

One possible measure to reduce such bias is through training. If exams are offered in both writing modes, it might be possible to train markers to ignore differences pertaining to each mode. However, there remain very limited studies on the effectiveness of training in reducing presentation mode effects on marker bias.

One important caveat to keep in mind regarding the literature on mode bias in marking, is that most of the relevant studies are at least 20 years old and took place before on-screen marking of scanned paper exam scripts became common practice. Given some of the possible contributors to bias relate to handwriting and legibility, which would be visible in scans of handwritten essays, there is still potential for there to be bias in current marking. On the other hand, markers' expectations of students' word-processed essays might have changed over time. Further evidence on whether bias against typed essays is present in current

© Cambridge University Press & Assessment 2024

marking, including when both handwritten and typed essays are marked on screen, would be valuable.

For **comparability of text characteristics**, the most frequent characteristic compared was text length. Word-processed essays tended to be longer than handwritten essays. However, the length of word-processed essays also appeared to vary more than that of handwritten essays. As computer familiarity could affect the length of essays produced, caution must be exercised to mitigate any risk of markers being biased by essay length. Although essay length has often been found to strongly correlate with scores and/or to be a strong predictor of scores (see Jeon & Strube, 2021; Kobrin et al., 2007), it is an irrelevant construct to writing. If Artificial Intelligence (e.g., an automated essay scoring system) is used for marking, it is crucial to ensure that the system does not rely on essay length in generating scores (see Jeon & Strube, 2021; Madnani & Cahill, 2018; Perelman, 2014). Using an automated scoring system that relies on construct-irrelevant features, including essay length, could threaten the interpretation of scores generated by the system (Bejar, 2017). Other differences in text characteristics such as language complexity and frequency and type of errors were also observed, but they were usually of little practical significance and may not necessarily translate to score differences.

There is a dearth of research examining **the comparability of composing processes** under the two writing modes. The few existing studies indicated that both modes elicit comparable processes with some minor differences. Comparable composing processes imply that both writing modes activate similar cognitive processes from students while they are engaged in task completion. Establishing cognitive equivalence between modes of composition becomes crucial when both modes are made available and schools may choose a composition mode on which their students are going to take the test.

In conclusion, potential mode effects due to writing mode can generally be considered a mature field of inquiry, evidenced by the number of empirical studies included in this review. Variability in research contexts, focuses and methods also further evidences the maturity of the research area. Such variability partly explains the differences in findings presented in this article. It should also be noted that some studies included in this review were conducted quite a while ago. Therefore, the generalisability and applicability of the findings should be considered carefully, given that both students and markers are likely to have increased familiarity and comfort with using a computer. An important aspect of writing mode effects in exams that remains little explored is the congruence between mode of learning and mode of testing and the extent to which this could influence mode effects.

Acknowledgement

The author would like to thank Camilo Ramos for earlier discussions of the literature in this area.

© Cambridge University Press & Assessment 2024

References

Arce-Ferrer, A. J., & Bulut, O. (2018). Effects of data-collection designs in the comparison of computer-based and paper-based tests. *The Journal of Experimental Education, 87*(4), 661–679.

Arnold, V., Legas, J., Obler, S., Pacheco, M. A., Russell, C., & Umbdenstock, L. (1990). *Do students get higher scores on their word-processed papers? A study of bias in scoring hand-written vs. word-processed papers*. Rio Hondo College.

Barkaoui, K., & Knouzi, I. (2018). The effects of writing mode and computer ability on L2 test-takers' essay characteristics and scores. *Assessing Writing*, *36*, 19–31.

Bejar, I. I. (2017). Threats to score meaning in automated scoring. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 75–84). Routledge.

Breland, H., Lee, Y. W., & Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts: Response mode analyses* (Research Report, Issue RR-75). Educational Testing Service.

Bridgeman, B., & Cooper, P. (1998, April 13–17). *Comparability of scores on word-processed and handwritten essays on the Graduate Management Admissions Test* [Paper presentation]. The Annual Meeting of the American Educational Research Association, San Diego, CA, United States.

Brown, A. (2003). *Legibility and the rating of second language writing: An investigation of the rating of handwritten and word-processed IELTS Task Two essays* (IELTS Research Reports, Issue 4). IDP: IELTS Australia.

Brunfaut, T., Harding, L., & Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing*, *36*, 3–18.

Burke, J. N., & Cizek, G. J. (2006). Effects of composition mode and self-perceived computer skills on essay scores of sixth graders. *Assessing Writing*, *11*(3), 148–166.

Chambers, L. (2008). Computer-based and paper-based writing assessment: a comparative text analysis. *Research Notes*, *34*, 9–15.

Chan, S., Bax, S., & Weir, C. (2018). Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test. *Assessing Writing*, *36*, 32–48.

Charman, M. (2014). Linguistic analysis of extended examination answers: Differences between on-screen and paper-based, high- and low-scoring answers. *British Journal of Educational Technology*, *45*(5), 834–843.

© Cambridge University Press & Assessment 2024

Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing*, *16*(1), 49–71.

Connelly, V., Gee, D., & Walsh, E. (2007). A comparison of keyboarded and handwritten compositions and the relationship with transcription speed. *British Journal of Educational Psychology*, *77*(2), 479–492.

Coombe, G., Lester, A., & Moores, L. (2020). *Online and on-screen assessment in high-stakes, sessional qualifications: A review of the barriers to greater adoption and how these might be overcome*. (Ofqual/20/6723/1).

Endres, H. (2012). A comparability study of computer-based and paper-based Writing tests. *Research Notes*, *49*, 26–33.

Green, T., & Maycock, L. (2004). Computer-based IELTS and paper-based versions of IELTS. *Research Notes*, *18*, 3–6.

Hollenbeck, K., Tindal, G., & Almond, P. (1999). Reliability and decision consistency: An analysis of writing mode at two times on a statewide test. *Educational Assessment*, *6*(1), 23–40.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, *5*(2), n2.

Jeon, S., & Strube, M. (2021). Countering the influence of essay length in neural essay scoring. The second workshop on simple and efficient natural language processing. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, 32–38, Virtual. Association for Computational Linguistics.

Jin, Y., & Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly*, *14*(2), 101–119.

Kim, H. R., Bowles, M., Yan, X., & Chung, S. J. (2018). Examining the comparability between paper- and computer-based versions of an integrated writing placement test. *Assessing Writing*, *36*, 49–62.

Kobrin, J. L., Deng, H., & Shaw, E. J. (2007). Does quantity equal quality? The relationship between length of response and scores on the SAT essay. *Journal of Applied Testing Technology, 8*(1), 1–15.

Lam, F., & Pennington, M. C. (1995). The computer vs. the pen: A comparative study of word processing in a Hong Kong secondary classroom. *Computer Assisted Language Learning*, *8*(1), 75–92.

© Cambridge University Press & Assessment 2024

Lee, H. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing*, *9*(1), 4–26.

Lee, Y.-J. (2002). A comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes. *Assessing Writing*, *8*(2), 135–157.

Lessien, E. (2013). *The effects of typed versus handwritten essays on students' scores on proficiency tests* [Unpublished doctoral dissertation, Michigan State University].

Lovett, B. J., Lewandowski, L. J., Berger, C., & Gathje, R. A. (2010). Effects of response mode and time allotment on college students' writing. *Journal of College Reading and Learning*, *40*(2), 64–79.

MacCann, R., Eastment, B., & Pickering, S. (2002). Responding to free response examination questions: Computer versus pen and paper. *British Journal of Educational Technology*, *33*(2), 173–188.

Madnani, N. & Cahill, A. (2018). Automated scoring: Beyond natural language processing. *Proceedings of the 27th International Conference on Computational Linguistics,* 1099–1109.

Manalo, J. R., & Wolfe, E. W. (2000a, April 24–28). *A comparison of word-processed and handwritten essays written for the Test of English as a Foreign Language* [Paper presentation]. The Annual Meeting of the American Educational Research Association, New Orleans, LA, United States.

Manalo, J. R., & Wolfe, E. W. (2000b, April 24–28). *The impact of composition medium on essay raters in foreign language testing* [Paper presentation]. The Annual Meeting of the American Educational Research Association, New Orleans, LA, United States.

McGuire, D. W. (1995). *A comparison of scores on the Kansas Writing Assessment for word-processed and hand-written papers of eleventh graders* [Unpublished doctoral dissertation, Kansas State University].

Mogey, N., & Hartley, J. (2013). To write or to type? The effects of handwriting and word-processing on the written style of examination essays. *Innovations in Education and Teaching International*, *50*(1), 85–93.

Mogey, N., Paterson, J., Burk, J., & Purcell, M. (2010). Typing compared with handwriting for essay examinations at university: Letting the students choose. *Research in Learning Technology*, *18*(1), 29–47.

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing*, *21*, 104–111.

© Cambridge University Press & Assessment 2024

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement*, *31*(3), 220–233.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, *5*(3).

Russell, M., & Plati, T. (2002). Does it matter with what I write? Comparing performance on paper, computer and portable writing devices. *Current Issues in Education*, *5*.

Russell, M., & Tao, W. (2004a). Effects of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment, Research, and Evaluation*, *9*(1).

Russell, M., & Tao, W. (2004b). The influence of computer-print on rater scores. *Practical Assessment, Research, and Evaluation*, *9*(1).

Russell, M. K. (1999). *Testing on computers: A follow-up study comparing performance on computer and on paper* [Unpublished doctoral dissertation, Boston College].

Shaw, S. D. (2003). Legibility and the rating of second language writing: The effect on examiners when assessing handwritten and word-processed scripts. *Research Notes*, *11*, 7–11.

Sweedler-Brown, C. O. (1991). Computers and assessment: The effect of typing versus handwriting on the holistic scoring of essays. *Research and Teaching in Developmental Education*, *8*(1), 5–14.

Whithaus, C., Harrison, S. B., & Midyette, J. (2008). Keyboarding compared with handwriting on a high-stakes writing assessment: Student choice of composing medium, raters' perceptions, and text quality. *Assessing Writing*, *13*(1), 4–25.

Wolfe, E. W., Bolton, S., Feltovich, B., & Bangert, A. W. (1995, April 18–22). *The influence of computers on student performance on a direct writing assessment* [Paper presentation]. The Annual Meeting of the American Educational Research Association, San Francisco, CA, United States.

Wolfe, E. W., Bolton, S., Feltovich, B., & Niday, D. M. (1996). The influence of student experience with word processors on the quality of essays written for a direct writing assessment. *Assessing Writing*, *3*(2), 123–147.

Wolfe, E. W., Bolton, S., Feltovich, B., & Welch, C. (1993). *A comparison of word-processed and handwritten essays from a standardized writing assessment* (ACT Research Report Series, Issue 93-8).

© Cambridge University Press & Assessment 2024

Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning & Technology*, *8*(1), 53–65.

Wolfe, E. W., & Manalo, J. R. (2005). *An investigation of the impact of composition medium on the quality of TOEFL Writing scores* (Research Report, Issue RR-72). Educational Testing Service.

Yu, L., Livingston, S. A., Larkin, K. C., & Bonett, J. (2004). *Investigating differences in examinee performance between computer-based and handwritten essays* (Research Report, Issue RR-04-18). Educational Testing Service.

Yu, W., & Iwashita, N. (2021). Comparison of test performance on paper-based testing (PBT) and computer-based testing (CBT) by English-majored undergraduate students in China. *Language Testing in Asia*, *11*(1), 32.

Zhu, Y., Shum, S.-K. M., Tse, S.-K. B., & Liu, J. J. (2016). Word-processor or pencil-and-paper? A comparison of students' writing in Chinese as a foreign language. *Computer Assisted Language Learning*, *29*(3), 596–617.

© Cambridge University Press & Assessment 2024